学校编码：10384　　分类号_____密级_

学号：X2008230120　　UDC_

厦门大学

硕 士 学 位 论 文

决策树理论在保险行业绩效评价中的应用研究

Research on Performance Evaluation of Insurance Industry Based on Decision Tree

龚鸿乾

指导教师姓名：陈海山　教授

专 业 名 称：软 件 工 程

论文提交日期：2010 年 10 月

论文答辩时间：2010 年　　月

学位授予日期：　　年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2010 年 10 月

摘 要

数据挖掘技术可以帮助人们从数据库，特别是数据仓库的相关数据集中提取出所感兴趣的知识、规律或更高层次的信息，并可以协助人们从不同程度上去分析它们，从而更有效地利用数据库或数据仓库中的数据。数据挖掘技术不仅可以用于描述过去数据的发展过程，进一步还能够预测未来趋势。预测分类作为数据挖掘的一项重要任务，在医疗诊断，气象预报，信贷审核，等诸多领域有着广泛应用。

多年来，保险公司在对员工的培训和管理工作中积累了大量的数据，目前这些数据还未能得到有效的利用，只是一个待开发的“宝藏”。鉴于公司对员工素质提高的需求和目前公司数据管理的现状，利用这些数据更合理地分析公司各方面工作的成效变得十分重要。

本文主要以笔者所在的单位为例，针对公司面临的职工人数增加、培训压力增大等实际问题，提出方案来提高员工培训质量。本文主要采用数据挖掘技术中的决策树方法，将其应用到员工的绩效评价中，利用收集到的大量与绩效有关的数据信息，转换成知识，以便指导公司在今后的培训工作中采取恰当的方法，指导员工工作，进而提高公司的经济效益。

关键词: 决策树；C4.5 算法；绩效评价

Abstract

The technology of data mining can help us to find the knowledge, rules or information of high level from the related data in the Database, especially the Data Warehouse. Classified discovery is widely used in many fields such as medical treatment, weather prediction, and credit verification.

Over the years, the insurance company has accumulated a large amount of data in training and managing the employees. The data is not used effectively at present, but it is an undeveloped "treasure" indeed. In consideration of the company' needs to improve the quality of employees and the current status of data management; it has become very important to analyze the effectiveness of the work in all aspects with the data.

For the problems of the grower number of the employees and the increasing pressure for staff training and others which insurance company faces, this paper proposes a solution to improve the quality of staff training in the case of author's company. By use of transforming large quantities of performance-related data information to knowledge, this paper uses data mining technology, mainly the decision tree method, to evaluate the employee performance, and to lead the company to adopt appropriate methods to guide staff work in training so as to improve the company's economic efficiency.

Keywords: Decision Tree; C4.5 Algorithm; Performance Evaluation

目　录

Contents

第一章　绪论

1.1 课题研究的背景和意义

数据挖掘技术在商业、金融业、保险业、市场营销等诸多领域已获得了较为广泛的应用，但对职工绩效信息进行数据挖掘与知识发现的研究和应用相对来说较少。公司对员工绩效信息等数据的处理还停留在简单的数据备份和查询阶段。

近年来，保险行业的业务的不断扩大，员工人数大幅度增加，这种新变化给公司员工管理及指导工作带来了严峻的考验，传统的管理手段已经逐渐不能适应社会的发展。随着数据挖掘技术的成熟及应用领域的不断扩展，不少公司已开始研究将数据挖掘技术应用于公司员工培训管理中。

公司的发展壮大的重点和关键是提高整个公司员工的工作素质，而员工绩效恰恰是评估员工工作质量的重要依据，也是评价员工对所学行业知识掌握程度的重要标志。所以通过对员工的绩效进行正确科学的分析评估，可以为引导公司管理层重视培训工作，注意改善培训条件，加强培训管理，深化培训改革，努力提高培训质量等提供重要的依据。

数据挖掘是一种决策支持过程，是深层次的数据信息分析方法，将数据挖掘技术应用于绩效评估方面是非常有益的，它可以全面地分析考核绩效与各种因素之间隐藏的内在联系，比如，经过对员工相关数据进行分析，数据挖掘工具可以回答诸如“哪些因素对员工绩效可能有影响”等类似的问题，这是传统评价方法无法具备的。

利用数据挖掘工具，对员工的绩效数据进行分析处理，可以及时得到员工的评价结果，对员工出现的不良工作行为进行及时指正。另外，还能够克服公司主观评价的不公正、不客观等弱点，并能够减轻公司管理层的工作量。绩效作为考核的结果，不仅是对员工业绩和公司培训及管理效果的检查和评定，进而提高员工学习及工作的效率；它更是一种信息，具有反馈于培训活动、服务于商业决策、为公司提供资料等作用。为充分发挥考核的效能，综合评价管理质量，及时反馈培训效果，沟通培训信息，指导部门对考核绩效进行统计分析和总结是非常必要的。

1.2 课题研究的现状

经过十多年的发展，数据挖掘技术的研究已经取得了丰硕的成果。目前，数据挖掘应用于许多领域，尤其在电信、银行保险、零售等商业领域。数据挖掘所能解决的典型问题包括：数据库营销、客户群体划分、交叉销售、背景销售等市场分析行为、客户信用积分、客户流失性分析、客户信用积分、欺诈识别等。

但国内关于如何将数据挖掘技术应用于保险业员工绩效分析这方面的研究相对来说还是空白，国内保险行业还是运用传统的统计学方法，并不能有效的进行规律的总结和知识的发现。

1.3 论文的研究内容及结构安排

公司多年来对员工的培训和管理工作中积累了大量的数据，目前这些数据还未能得到有效的利用，只是一个待开发的“宝藏”。鉴于公司对员工素质提高的需求和目前公司数据管理的现状，利用这些数据理性地分析公司各方面工作的成效以及员工培训过程中的得失变得十分重要。

本文主要以笔者所在的单位为例，针对公司面临的职工人数增加、培训压力增大等实际问题，提出方案来提高员工培训质量。本文主要采用数据挖掘技术中的决策树方法，将其应用到员工的绩效分析中，利用收集到的大量与绩效有关的数据信息，并将这些信息转换成知识，以便指导公司在以后的培训工作中采取恰当的方法，指导员工的工作，进而提高公司的经济效益，是本文探讨的主要内容。

全文的各章结构安排如下：

第一章——绪论。主要介绍本课题提出的背景和意义，并提出本论文所研究的内容及论文结构的安排。

第二章——数据挖掘与知识发现。系统的介绍数据挖掘几个重要的概念，以及其中几个较为突出的算法，并讨论了进行数据挖掘的基本步骤，为课题的展开做理论铺垫。

第三章——决策树技术。分析数据挖掘中分类技术的概念、决策树建立的过程及选择属性的量度、决策树中的剪枝技术，并详细分析决策树技术中主要的算法。

第四章——对决策树技术在绩效分析中的应用的研究。这章是本论文的重点，详细并完整的实现决策树挖掘技术在绩效分析中的应用全过程，最后建立了两个分析模型，

包括员工绩效是否优良的决策树模型和绩效是否不及格的决策树模型。

第五章——保险业员工绩效分析系统的设计及实现。基于上述章节的理论和知识铺垫，利用高级程序设计语言 Delphi7，结合数据库开发技术，实现一个完整的绩效分析系统。

第六章——总结与展望。主要总结了在论文研究中所进行的工作，并对存在的不足进行了分析，并对日后的工作做了规划。

第二章　数据挖掘与知识发现

数据挖掘(Data Mining)是一个多学科交叉研究领域，它融合了数据库（Database 技术、人工智能（Artificial Intelligence)、机器学习（Machine Learning)、统计学（Statistics）知识工程（Knowledge Engineering)、面向对象程序设计（Object-Oriented Method)、信息检索（Information Retrieval)、高性能计算（High-Performance）等最新技术的研究成果。经过十几年的研究，产生了许多新概念和方法。特别是最近几年，一些基本概念和方法趋于清晰，它的研究正向着更深入的方向发展。数据挖掘之所以被成为未来信息处理的骨干技术之一，主要在于它以一种全新的概念改变着人类利用数据的方式。二十世纪，数据库技术取得了决定性的成果并且已经得到了广泛的应用。但是，数据库技术作为一种基本的信息存储和管理方式，仍然以联机事务处理（OLTP: On-Line Transaction Processing）为核心应用，缺少对决策、分析、预测等高级功能的支持机制。众所周知随着数据库容量的膨胀，特别是数据仓库（Data Warehouse）以及 Web 等新型数据源的日益普及，联机分析处理（OLAP: On-Line Analytic Processing)、决策支持（Decision Support）以及分类（Classing）[1]、聚类（Clustering）[2]等复杂应用成为必然。面对这一挑战，数据挖掘和知识发现（Knowledge Discovery）技术应运而生，并显示出强大的生命力。数据挖掘和知识发现使数据处理技术进入了一个更高级的阶段。它不仅能对过去的数据进行查询，并且能够找出过去数据之间的潜在联系，进行更高层次的分析以便更好地做出理想的决策、预测未来的发展趋势。

从数据库中发现知识(Knowledge Discovery in Database，KDD)是 20 世纪 80 年代末开始的，KDD 一词是在 1989 年 8 月十美国底特律市合并的第一届 KDD 国际学术会议上正式形成的。KDD 研究的问题有：(1)定性知识和定量知识的发现；(2)知识发现方法；(3)知识发现的应用等 [3]。

1995 年在加拿大召开了第一届知识发现和数据挖掘(Data Mining，DM)国际学术会议 [4]。由于把数据库中的“数据”形象地比喻成矿床，“数据挖掘”一词很快流传开来。

数据挖掘是知识发现中的核心工作，主要研究发现知识的各种方法和技术。数据挖掘首先要确定挖掘的任务或目的。确定了挖掘任务后，就要决定使用的挖掘算法。选择实现算法有两个考虑因素：一是不同的数据有不同的特点，因此需要用与之相关的算法来挖掘，二是用户或实际运行系统的要求。选择了挖掘算法后，就可以实施数据挖掘操

作，获取有用的模式。

数据挖掘作为知识发现过程的一个特定步骤，它是一系列技术及应用，或者说是对大容量数据及数据间关系进行考察和建模的方法集。它的目标是将大容量数据转化为有用的知识和信息。

一般情况下，数据挖掘的对象定义为数据库，而更广义的说法是，数据挖掘意味着从一些实事或观察数据的集合中寻找模式。数据挖掘的对象不仅是数据库，也可以是文件系统或者其他任何组织在一起的数据集合，例如 Internet 信息资源、数据仓库等。数据挖掘广义定义：数据挖掘是从存放在数据库、数据仓库或其它信息库中的大量数据中挖掘有趣知识的过程[5]。

与数据挖掘和知识发现关系密切的研究领域包括归纳学习（Inductive Learning）机器学习（Machine Learning）和统计（Statistics）分析，特别是机器学习被认为和数据挖掘的关系最密切[6][7]。

数据挖掘的技术基础就是人工智能，它利用了人工智能中的诸多算法进行挖掘，为用户提供有用信息[8]。在很大程度上，数据挖掘是人工智能的某些成熟的技术(人工神经网络、遗传算法、决策树)在特定的应用系统中具体而微小的应用，但是其问题的规模和难度大大降低。

除了人工智能之外，数据挖掘还结合了传统的统计分析、模糊数学以及科学计算可视化技术，以数据库和数据仓库为研究对象，形成了数据挖掘方法和技术[9][10][11]。

在现实生活中，数据挖掘技术被广泛应用，尤其是在决策支持系统中，常常利用数据挖掘技术从数据库系统中获得有用信息。让更高一级的用户根据挖掘结果做出更明智、更正确的决策。

2.1 数据挖掘的概念

2.1.1 数据挖掘的定义

数据挖掘（Data Mining）[12]就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程，与数据挖掘相近的同义词有数据融合、数据分析和决策支持等。这个定义包括好几层含义：数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要求可接受、可理解、可运用；并不要求发现放之四海皆准的知识仅仅支持

特定的发现问题[16-19]。

与传统信息处理方法相比，数据挖掘技术有其自身的特点。

1、处理对象为大规模数据库，数据规模十分巨大，待处理数据的规模可能达到GBHB 甚至更大；

2、信息查询一般是由决策制定者（用户）提出的即时随即查询，往往没有更精确的查询要求，需要靠数据挖掘技术寻找其可能感兴趣的东西；

3、在一些应用中，某些行动并没有实际发生或很少发生，因而他们对输出所造成的影响没有在数据库中体现出来，需要利用数据挖掘技术从数据库中提取有用的规则为这种情况提出预测；

4、在一些应用中，由于数据变化迅速，因此要求数据挖掘技术能快速对数据变化作出反映以提供决策支持。数据挖掘既要发现潜在的规则，还要管理和维护规则，而规则是动态的，当前的规则只能反映当前状态的数据库特征，随着新数据的不断加入，规则随之更新；

5、数据挖掘中规则的发现主要基于大样本的统计规律，发现的规则不必适用于所有的数据，当达到某一阀值时，便可认为有此规律。

2.1.2 数据挖掘与知识发现的联系与区别

从应用深度上,我们将数据挖掘划分为三个层次空间。

1、数据空间:它利用现有数据库管理系统的查询和报表功能，进行基于关键字的决策查询，实现联机事务处理(On-Line Transaction Processing,简称 OLTP)。

2、聚合空间：利用聚集运算（Sum、Ave、Max、Min），结合多维分析和统计分析实现在线分析处理（On-Line Analytical Processing，简称 OLAP），以提供决策参考的统计分析数据。

3、影响空间：按照相似性的聚类、差异性的分类方法，发现关联性及结构模式、顺序模式、相似时序，建立预测模型，从数据库或大量数据记录中发现隐含的有用信息这是在更深层次上的知识发现，是数据挖掘实质性内涵。

以上数据挖掘的各个层次空间反应了不同级别的查询要求，这种划分有利于制止的逐步提取，知识的提取过程即为决策支持过程。

知识发现（KDD）是指存在于数据库中有效的、新颖的、具有潜在效用的、最终可

理解的、模式的，非平凡过程[20][21]。它是一个众多学科相互交融形成的、有广阔应用前景的新兴领域，其中包括人工智能、机器学习、模式识别、统计学、数据库及知识库等。知识发现（KDD）的整个过程包括在指定的数据库中用数据挖掘算法提取模型，以及围绕数据挖掘进行的预处理和结果表达等一系列计算步骤。数据挖掘算法位于整个过程的核心，通常占整个过程的 15%-25%的工作量。

数据挖掘（Data Mining）则是知识发现的核心环节。数据挖掘是知识发现的一个关键步骤，包括特定的数据挖掘算法，具有可接受的计算效率，生成特殊的模式；知识发现强调知识是数据发现的最终产品，利用响应的数据挖掘算法，按指定方式和阀值提取有价值的知识，包括数据挖掘对数据的预处理，抽样及转换和数据挖掘后对知识的评价解释过程。

2.1.3 数据挖掘的对象

根据信息存储格式，用语挖掘的对象有事务数据库、关系数据库、面向对象数据库数据仓库、文本数据源、多媒体数据库、遗产数据库以及 Web 页。

目前，用语数据挖掘的数据源主要是事务数据库、关系数据库、数据仓库和互联网 Web 页。

2.1.4 数据挖掘的相关办法

数据挖掘有多种方法。从数据挖掘的任务的角度可分为：关联规则挖掘、序列模式挖掘、聚类数据挖掘、分类数据挖掘、偏差分析挖掘和趋势预测挖掘等[22][23]。

分类数据挖掘也称为分类发现(Classification)是数据挖掘中一项非常重要的任务，在科学实验，医疗诊断，气象预报，信贷审核，商业预测，案件侦破等领域有着广泛应用。分类发现的目的是构造一个分类函数或分类模型，通过分类函数，把数据库中的元组映射到给定类别中的某一个。进而发现一些指定的事件是否属于某一特定数据子集的规则。

一般把分类模型中的输入称为“训练集”，它的每一个元组的属性和数据库的元组的属性相同，并且，每个元组都有一个已知的类别标志。分类的目标是通过分析训练集中的数据，发现个体或对象的一般分类规则，对类进行准确的描述或者建立模型，然后用它对数据库中的其它非样本数据进行分类。

依据分类发现采用的分类模型不同，有多种分类的方法，研究和应用最多的方法主

要有：基于决策树模型的数据分类，基于统计模型的数据分类，基于神经网络模型的数据分类。

2.1.5 数据挖掘的技术

数据挖掘中的关键技术是进行模式识别和关系识别的算法,许多算法来源于人工智能和机器学习等领域。

数据挖掘分两类:预言性数据挖掘和描述性数据挖掘。预言性数据挖掘是进行数据分析，建立一个或一组模型，并根据模型产生关于数据的预测：描述性数据挖掘是以概要的方式对数据信息进行描述，提供数据的对用户有趣的一般性质。

预言性数据挖掘分析：采用的主要方法是分类，分类是根据训练集数据找到可以描述并区分数据类别的分类模型，使之可以预测未知数据的类别。分类可以采用神经元网络算法和决策树算法。

描述性数据挖掘分析：包括异常监测、聚集等多种数据挖掘方法。异常检测是数据挖掘中一个重要方法，用来发现“小的模式”，即找到数据集中与大多数数据不同或不一致的数据对象。聚集则是把数据集分为不同的簇，使得簇与簇之间的差别明显，而簇内个体之间的差异较小。

2.2 数据挖掘的步骤

数据挖掘的特点之一是在真正开始数据挖掘之前需要做大量预处理工作。这些工作包括：定义问题、获取相关数据和为挖掘准备数据。

下面对数据挖掘的一般办法和步骤进行具体阐述。

2.2.1 定义问题

这一步的主要目的是确定数据挖掘是否适合解决客户所提出的问题，其次还要做两项相关工作：

第一，需要从客户那里获取哪些数据；

第二，是否有足够的数据支持数据挖掘。

2.2.2 获取数据

这一阶段主要是在数据库专家的帮助下理解数据库的结构、内容等。